Modality-Specific Training in Audio-Visual Speech Integration

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation *with distinction* in Speech and Hearing Science in the undergraduate colleges of The Ohio State University

By

Ashley Case

The Ohio State University
June 2012

Advisor: Dr. Janet M. Weisenberger, Department of Speech and Hearing Science

Abstract

Listeners integrate auditory and visual signals to understand speech in both compromised and normal listening environments. This integration appears to be a process independent of the ability to process auditory-only or visual- only speech cues (Grant & Seitz, 1998). Training with auditory-only stimuli does not seem to generalize to the audio-visual condition (James, 2009); nor does audio-visual training produce improvements in auditory-only perception. Because skill in all three modalities is important in speech perception by hearing impaired persons, the question remains whether audio-visual integration would improve if training in all three modalities were provided. In the present study, five listeners received ten training sessions that included auditory-only, visual-only and audio-visual stimuli. The auditory component of these speech stimuli was degraded in a similar method to Shannon et al. (1995). Results showed that subjects improved in all conditions; however, different measures of audio-visual integration yielded conflicting indications of integration improvement. These results suggest that stimulus selection plays an important role in training to improve audio-visual integration in aural rehabilitation programs.

Acknowledgements

I would like to thank my advisor, Dr. Janet M. Weisenberger, for her guidance and support throughout the entire process of research. I have learned a great deal from her, both within the discipline of my thesis and in general, and I am very thankful to have had this experience. I would also like to thank my subjects for their dedication and flexibility while conducting the research.

This project was supported by an ASC Undergraduate Scholarship and an SBS Undergraduate Research Scholarship.

Tables of Contents

Chapter 1: Introduction and Literature Review

Contrary to the belief that speech perception is strictly an auditory process, researchers have shown that listeners use information from both the auditory and visual signals to understand speech, especially when the auditory signal is compromised in some manner. In these situations, such as noisy environments or hearing loss, visual cues can help to fill in information that was missing in the compromised auditory signal. However, research by McGurk and MacDonald (1976) revealed that that even individuals provided with a perfect auditory signal could not ignore incoming visual information in speech perception.

McGurk and MacDonald (1976) demonstrated that visual cues are used even when the auditory signal is perfect. In their study, McGurk and MacDonald dubbed auditory syllables onto video recordings of different syllables in order to measure the degree to which audio-visual integration occurred. When listeners were presented with the auditory signal /ba/ and the visual signal /ga/, the brain integrated the two signals such that the participants perceived the intermediate phoneme /da/. This response was classified as a "fusion" due to the fact that the listener perceived a response different from either of the two inputs, presumably by fusing the bilabial position of the auditory /ba/ and the velar position of the visual /ga/. As a result, the listener perceived the intermediate syllable /da/, which lies in the alveolar place of articulation between the positions of the two presented signals. McGurk and MacDonald discovered that this type of response was not the only one possible in speech perception integration. By switching the auditory and visual stimuli to present an auditory /ga/ signal and visual /ba/ signal, McGurk and MacDonald found that listeners reported hearing a "combination" response of /bga/. In this instance, the bilabial syllable /ba/ presented strong visual information. The brain could not ignore the prominent information presented in the visual signal and therefore combined the two inputs,

rather than fusing them to create an entirely new syllable. In both cases, however, the listener's response indicated audio-visual integration, which came to be known as the McGurk effect. Additional studies have shown that audio-visual integration is an automatic behavior that occurs unconsciously among all listeners. In order to better understand the phenomenon of audio-visual integration, it is necessary to consider the auditory and visual cues that are available for listeners in speech perception.

Auditory Cues for Speech Perception

The auditory signal provides listeners with three main cues for consonant identification. These cues include place of articulation, manner of articulation and voicing. The first cue, place of articulation, refers to where in the vocal tract the sound is produced, or the point at which the oral cavity is obstructed during sound production. Possible places of articulation include bilabial, labiodental, dental, alveolar, palatal, velar, glottal and lingual. The second cue, manner of articulation, describes the method in which the airstream is modified as it moves through the vocal tract. Possible manners of articulation include stops, fricatives, affricates, nasals, liquids and glides. The third cue, voicing, refers to the presence or absence of vocal fold vibration during speech production. The presence of vocal fold vibration indicates a voiced consonant, whereas the absence of vocal fold vibration indicates a voiceless consonant. (Small, 2004). All the cues are represented in the acoustic signal, in formant transitions, resonant frequencies, turbulence, and voice onset time.

Visual Cues for Speech Perception

Although auditory cues provide a great deal of information about the speech sound that has been produced, McGurk and MacDonald presented evidence that visual cues are also

important in speech perception. Unlike the auditory signal that provides information regarding place, manner and voicing, the visual signal can only be a reliable source of information for the place of articulation (Jackson, 1988). Even place cues are not perfectly represented visually. For this reason, it may be difficult to identify a speech sound based solely on visual information.

The difficulty of identifying speech sounds solely on visual information is a result of groups that are referred to as visual phonemes, or visemes (Jackson, 1988). Speech sounds within a viseme group have identical visual features, but differ in manner and voicing. For example, the phonemes /p, b, m/ occur in the bilabial place of articulation and therefore constitute a viseme group. These sounds are not distinguishable with vision alone. There are five universal viseme groups which are most commonly identified and accepted as groupings because of their identical visual features. These are /p, b, m/, /f, v/, /θ, ð/, /ʃ, ʒ, tʒ, dʒ/, and /w, r/.

When there is no auditory signal to aid in speech perception, speechreading may also prove more difficult as a result of individual talker characteristics. A study by Jackson (1988) concluded that it was easier to speechread the talkers who demonstrated a greater number of viseme categories. Those talkers who were the easiest to understand visually demonstrated the five universal visemes. Cues provided by the talker such as gestures and movements of the eyes, head and mouth may also assist the listener in speech perception. These cues are particularly helpful in situations where the auditory signal is compromised or absent.

Speech Perception with Reduced Auditory and Visual Signals

Even when the auditory signal is compromised, speech may still be highly intelligible, partially due to the fact that the speech signal contains redundant information. A study by Shannon and colleagues (1995) revealed that acoustic speech waveforms contained more

information than is absolutely necessary for speech recognition. In this study, speech sounds were degraded by a method that removed varying amounts of spectral information and replaced it with band-limited noise. The temporal envelope and amplitude cues were preserved, allowing Shannon and his colleagues to examine what impact temporal cues play on speech perception. They found that speech was relatively intelligible when the temporal envelope was intact, although the clarity of speech improved as the number of bands in the spectral cue increased. Yet even when nearly all spectral cues were removed, speech phoneme discrimination and recognition were still at a surprisingly high level. Shannon et al. found that because of this redundant information, listeners were able to identify speech sounds with as little as three to four modulated bands of noise.

In 1998, Shannon and his colleagues expanded this study to analyze the importance of spectral parameters in speech pattern recognition. Shannon et al. conducted four experiments in this study, including varying the location of band divisions, warping the spectral distribution of envelope cues, shifting the frequency of envelope cues in a tonotopic manner, and spectral smearing. Results showed that the exact cutoff frequencies that define the four modulated bands of noise and the selectivity of the envelope carrier bands were not critical for speech recognition. In contrast, warping the spectral distribution of envelope cues and shifting the tonotopic envelope pattern resulted in poor intelligibility of speech.

A study by Remez and colleagues (1981) also examined speech intelligibility under conditions in which the auditory signal was degraded. In this study, the auditory signal was transformed into a three-tone sine wave replica to represent the first three formants of the original signal. Although listeners recognized that the auditory signal provided to them was unnatural, the linguistic content within the signal was highly intelligible. This research, similar to

that conducted by Shannon et al. (1995), demonstrates that speech may be intelligible under degraded auditory conditions due to its redundant nature.

As seen in studies conducted by Shannon et al. and Remez et al., auditory information does not need to be perfect to aid in speech intelligibility. Similarly, degraded visual cues may make signals more intelligible for listeners. In a study by Munhall and colleagues (2004), visual images were degraded using band-pass and low-pass spatial filters. Auditory signals in noise were dubbed onto the degraded visual images and presented to the listeners. Results showed that information in a range of spatial frequencies enhanced auditory intelligibility, especially in the mid-range band. It was concluded that high spatial frequency information is not crucial in the visual cue for speech perception. Similar to the auditory signal, visual images do not need to be perfect to aid in speech perception.

Audio-Visual Integration of Reduced Information Stimuli

A study conducted in our laboratory by Feleppelle (2008) presented listeners with a degraded auditory signal similar to the signal which hearing-impaired individuals perceive. This study specifically examined whether the amount of information present in the auditory signal affected the degree to which audio-visual integration occurred in listeners. Speech perception abilities of listeners were observed in auditory-only (A), visual-only (V), and audio-visual (A+V) conditions. In order to reduce the information present in the auditory signal, Feleppelle removed varying amounts of spectral information and replaced it with noise, while preserving the temporal envelope. This signal degradation is comparable to that used by Shannon et al. (1998). Feleppelle tested subjects with four levels of this auditory degradation, using 2, 4, 6 and 8 band-pass filter channels. Results showed that although removing information from the auditory signal

negatively affected speech perception in auditory-only (A) and audio-visual (A+V) conditions, it did not inhibit integration entirely.

A study by Grant and Seitz (1998) evaluated how hearing-impaired persons integrate auditory and visual speech signals. In this study, subjects were presented with "congruent" and "discrepant" nonsense syllables in noise. Congruent syllables contain matching auditory and visual information (e.g., visual *gat* – auditory *gat*), whereas discrepant syllables contain auditory and visual information that do not match (e.g., visual *gat* – auditory *bat*). These degraded syllables were presented to listeners in three conditions: auditory (A), visual (V), and audio-visual (A+V). Results of this study showed that audio-visual benefit was very high, even in the case of an extremely compromised auditory signal. Yet because audio-visual integration could not be predicted from an individual's auditory-only and visual-only performance, Grant argued that audio-visual integration is an independent and measurable process from single-modality processing.

Effects of Training in Recent Studies

Various studies from our laboratory provide some support for the argument that audio-visual integration is an independent process as posited by Grant and Seitz (1998). Gariety (2009) and James (2009) examined how training with auditory-only presentation influenced audio-visual integration. Training in the auditory-only condition with degraded speech sounds improved only the listener's ability to perceive the auditory signal, but did not generalize to improve audio-visual integration.

Ranta (2010) and DiStefano (2010) addressed the question of whether audio-visual integration itself could be trained. While DiStefano demonstrated that integration could be

trained for "congruent" stimuli (i.e., both the auditory and visual signals were the same syllable), Ranta extended these findings to show that "discrepant" stimuli (i.e., auditory and visual signals were different syllables designed to evoke McGurk-type responses) could also be trained. For both studies, in which training was provided only in audio-visual integration, only integration improved; the single modality presentations of auditory-only and visual-only speech signals did not show improvement.

Because it is important to maximize skill in auditory, visual, and audio-visual speech perception in aural rehabilitation training, the question remains whether integration would improve if training in all three modalities were provided.

Present Study

The present study examined this question by training listeners in all three conditions: auditory-only, visual-only and audio-visual. Similar to the methods used by DiStefano (2010) and Ranta (2010), five listeners each received twelve hours of training. Participants were tested pre-training, mid-training and post-training. In contrast to the studies by DiStefano and Ranta, listeners in the present study were trained in auditory-only, visual-only and audio-visual conditions. It was anticipated that improvements would be seen in all conditions, and that integration would increase substantially after training. These data should provide insights into generalization of training with auditory and visual speech inputs and provide professionals in the field with a better understanding of the processes underlying auditory-only, visual-only and audio-visual speech perception. This knowledge will provide valuable information for the design of aural rehabilitation programs for individuals with hearing impairments.

Chapter 2: Method

Participants

Five listeners participated in the present study. Of the five listeners, two were female and three were male. Ages of the participants ranged from 18 to 22. All five were native American English speakers and reported having normal hearing, as well as normal or corrected vision. None of the participants had a background in Speech and Hearing Science. Participants were compensated approximately $10 for every hour of participation, totaling $160. Materials previously recorded by three talkers, 1 male, 2 female, all native speakers of American English, were used as stimuli.

Stimuli Selection

A closed set of eight syllables were presented in the study. All syllables presented had the following conditions:

1. The stimuli were “minimal pairs”, differing only in the initial consonant.
2. All syllables contained the vowel /ae/, chosen because it does not involve lip rounding or lip extension, which may create difficulties in lip reading.
3. Syllables in the stimulus set represented all possible categories of articulation, including place (bilabial, alveolar, velar), manner (stop, fricative, nasal) and voicing (voiced, unvoiced).
4. All were presented in isolation without a carrier phrase.

Stimuli

For each condition, the same set of single-syllable stimuli was used:

Bilabial: bat, mat, pat

Alveolar: sat, tat, zat

Velar: cat, gat

The following dual-syllable (dubbed) stimuli were also used in the degraded audio-visual condition. These were classified as "discrepant" stimuli. The first syllable represents the visual stimulus and the second syllable represents the auditory stimulus.

bat-gat

gat-bat

pat-cat

cat-pat

Stimuli Recording and Editing

Stimuli from recent studies (i.e., DiStefano, 2010; Ranta, 2010) were used in the study to permit comparison of results. To create these stimuli, speech samples from five talkers were degraded using a MATLAB script designed by Delgutte (2003). The speech signal was first filtered into two broad spectral bands. Then the fine structure component of each band was replaced with band-limited noise, while keeping the temporal envelope intact. The result was a 2-channel stimulus, similar to those used by Shannon et al. (1998). The degraded auditory stimuli were then dubbed onto the visual stimuli using Video Explosion Deluxe, a commercial video-editing program. Finally, the software program Sonic MY DVD was used to burn the stimulus sets onto DVDs. Four DVDs were created for each talker. Each DVD contained sixty stimuli,

organized in a random order to eliminate the possibility of memorization. DVDs for three of the five talkers were randomly selected as stimuli for the present study. For the visual-only presentation, the volume on the DVD player was set to "mute".

Visual Presentation

For the visual and audio-visual conditions, a 50 cm video monitor was positioned approximately 60 cm outside the window of a sound attenuating booth. The monitor was positioned at eye level, about 120 cm away from the participant seated inside the booth. Stimuli were presented using recorded DVDs on a DVD player. For auditory-only presentation the monitor screen was darkened.

Degraded Auditory Presentation

The degraded auditory stimuli were presented from the headphone output of the DVD player through 300-ohn TDH-39 headphones at the level of approximately 75 dBSPL. For visual-only presentation the volume on the DVD player was set to mute.

Testing Procedure

Testing was conducted in The Ohio State University's Speech and Hearing Department, located in Pressey Hall. Participants were instructed to read over a set of instructions explaining the procedure and listing a closed-set of response possibilities. The response set included more response possibilities than just the presented stimuli; it also included options that might reflect McGurk-type fusion or combination responses for the discrepant stimuli. The additional response choices included the syllables dat, nat, pcat, ptat, bgat and bdat.

Participants were individually tested in a sound-attenuating booth facing a video monitor outside the booth. Auditory stimuli were transmitted through headphones to the participants inside the booth. The examiner recorded and scored the participant's verbal responses through an intercom system. Each participant was administered a pre-test using stimuli selected from a set of 9 DVDs, each containing 60 randomly ordered syllables, three DVDs for each of the three talkers. In the pretest, the listeners were presented with one DVD from each talker in each of the three conditions (A, V, and A+V). Each DVD in the audio-visual condition included 30 stimuli with congruent auditory and visual components. The other 30 stimuli were discrepant, intended to elicit McGurk-type responses. Participants were asked to listen/watch each DVD and verbally respond with what they perceived. No feedback was provided during the pre-test.

The pretest was followed by five training sessions. Each session included training in the auditory-only, visual-only and auditory-visual conditions. DVDs were selected at random, using three DVDs from each of the three talkers, totaling nine DVDs for each session. The specific DVD and condition being trained were randomly selected. Feedback was provided during training in all conditions. In the auditory-only condition, visual-only condition, and audio-visual condition with congruent stimuli, the examiner verbally provided the correct response through the intercom. If the participant provided the correct answer, the examiner visually reinforced the participant with a head nod. For discrepant stimuli in the audio-visual condition, the appropriate McGurk-type component feedback was given to participants as the correct response. The feedback was given as follows, with the first syllable representing the visual stimulus, second syllable representing the auditory stimulus, and the third syllable representing the McGurk-type feedback provided:

bat-gat bgat

gat-bat dat

pat-cat pcat

cat-pat tat

A mid-test, using the same procedure as the pre-test, was administered following the first five training sessions. No feedback was provided. Five more training sessions including the auditory-only, visual-only and audio-visual conditions, with the same procedure as the first five training sessions, followed the mid-test. Finally, a post-test was conducted. No feedback was provided. Each test took approximately 2-3 hours, and the 10 training sessions took approximately 12-15 hours. Training sessions were broken up into one or two sessions at a time. Participants were encouraged to take frequent breaks in order to prevent fatigue.

Chapter 3: Results and Discussion

Percent Correct Performance

Figure 1 shows the overall percent-correct performance for pre-, mid- and post-tests, averaged across talkers and listeners. Performance in the auditory-only condition improved from 37% on the pre-test to 62% on the post-test. Performance in the visual-only condition improved from 34% on the pre-test to 51% on the post-test. Performance in the audio-visual conditions improved from 60% on the pre-test to 74% on the post-test. These results match those anticipated; listeners improved in all the conditions in which they were trained. A two-factor, repeated-measures ANOVA showed a significant main effect of test ($F(1,4)=20.850$, $p=.01$) and a significant main effect of modality ($F(2,8)=19.468$, $p=.001$). However, a significant interaction effect between test and modality was not observed ($F(2,8)=1.192$, ns).

The next several figures show performance for individual listeners. Figure 2 shows the percent-correct responses across tests for listener 1. Performance in the auditory-only condition showed improvement, increasing from 37% on the pre-test to 72% on the post-test. Performance in the visual-only and audio-visual conditions improved as well, but to lesser degrees. Performance in the visual condition increased from 33% on the pre-test to 44% on the post-test. Performance in the audio-visual condition increased from 72% on the pre-test to 78% on the post-test.

Figure 3 shows the percent-correct responses across tests for listener 2. Considerable improvements were seen for performance in the auditory-only and visual-only conditions, but performance in the A+V condition showed only minor improvement with training. Performance in the auditory-only condition improved from 34% on the pre-test to 74% on the post-test.

Performance in the visual-only condition improved from 22% on the pre-test to 72% on the post-test. Listener 2 showed a dramatic increase in visual-only performance in comparison to the other listeners, possibly because her performance was initially much poorer than that of the other listeners. Performance in the audio-visual condition increased from 68% on the pre-test to 80% on the post-test.

Figure 4 shows the percent-correct responses across tests for listener 3. Performance in the auditory-only condition increased from 44% on the pre-test to 57% on the post-test. Performance in the visual-only condition increased from 39% on the pre-test to 47% on the post-test. Performance in the audio-visual condition increased from 61% on the pre-test to 82% on the post-test.

Figure 5 shows the percent-correct responses across tests for listener 4. Performance in the auditory-only condition increased from 38% on the pre-test to 56% on the post-test. Performance in the visual-only condition increased from 33% on the pre-test to 46% on the post-test. Performance in the audio-visual condition increased from 64% on the pre-test to 72% on the post-test.

Figure 6 shows the percent-correct responses across test for listener 5. Performance in the auditory-only condition improved from 33% on the pre-test to 53% on the post-test. Performance in the visual-only condition remained at 43% from pre-test to post-test. Improvement in the audio-visual condition was dramatically better than the improvement seen in the single modalities. Performance improved from 36% on the pre-test to 57% on the post-test, again possibly because of much poorer performance on the pre-test compared to the other listeners.

Results were also examined in view of the intelligibility of individual talkers. Figure 7 shows the percent-correct responses across tests for talker 1, LG. Listeners performed the best in response to this talker's auditory and visual cues, in comparison to the other talkers, making LG the most intelligible talker. Performance in the auditory-only condition improved from 53% on the pre-test to 78% on the post-test. Performance in the visual-only condition improved from 37% on the pre-test to 60% on the post-test. Performance in the audio-visual condition improved from 67% to 82%.

Figure 8 shows the percent-correct responses across tests for talker 2, EA. Listeners performed well in response to this talker, but not as well as they did with talker LG. Performance in the auditory-only condition increased from 28% on the pre-test to 54% on the post-test. Performance in the visual-only condition increased from 33% on the pre-test to 45% on the post-test. Performance in the audio-visual condition improved from 57% on the pre-test to 74% on the post-test.

Figure 9 shows the percent-correct responses across tests for talker 3, JK. Generally listeners performed the worst with this talker, even with training. Performance in the auditory-only condition improved from 31% on the pre-test to 55% on the post-test. Performance in the visual-only condition improved from 33% on the pre-test to 47% on the post-test. Performance in the audio-visual condition improved from 57% on the pre-test to 65% on post-test.

Figure 10 shows the amount of audio-visual integration for the percent-correct stimuli across tests, by listener. Integration was defined as the difference between audio-visual performance and the better single modality, either auditory-only or visual-only. With this measure, results showed a decrease in integration from pre-test to post-test for listeners 1, 2 and

4. A decrease in integration was seen for all listeners from mid-test to post-test. This suggests that listeners were improving to a greater degree in the single modalities in comparison to audio-visual condition.

Responses to Discrepant Stimuli

Figure 11 shows the percent response scores across tests and listeners for discrepant, McGurk-type stimuli. For this figure, responses were broken up into "visual", "auditory" and "other" categories. "Visual" responses were those in which the listener replied with the visual component of the discrepant stimuli. Similarly, if listeners replied with the auditory component of the stimulus, this would be an "auditory" response. Finally, if listeners replied with a syllable other than the visual or auditory component of the stimulus, their response was placed into the "other" category. Looking at the figure, one can see that listeners decreased in the percentage of "visual" responses with training, dropping from 76% on the pre-test to 55% on the post-test. This led to a corresponding increase in the amount of "other" responses, increasing from 21% on the pre-test to 42% on the post-test. This increasing amount of "other" responses may indicate that the occurrence of integration. Further analysis of the "other" responses would be necessary before any conclusion could be made. The percentage of "auditory" responses was minimal and only increased from 2% to 3% across tests. The lack of reliance on the auditory signal was interesting given the increases in performance for auditory-only presentation with training, as shown in Figure 1. A paired samples t-test, ($t(4)=2.947$, $p=.04$), showed a significant change in the amount of "other" responses from pre-test to post-test averaged across listeners and tests.

Figure 12 shows further analysis of the "other" responses from figure 11. Listeners made more fusion McGurk-type responses with training, increasing from 36% on the pre-test to 51%

on the post-test. Listeners also made more combination McGurk-type responses with training, increasing from 26% on the pre-test to 45% on the post-test. The increase in both categories of McGurk-type responses indicates that integration was occurring for discrepant, McGurk-type stimuli, and that integration increased with training. The decrease in "neither" responses, those which were not fusion or combination McGurk responses, supports the occurrence of integration. A paired samples t-test shows a significant decrease in the percentage of "neither" responses, ($t(4)=4.548$, $p=.01$). This finding is in contrast to the observed decrease in integration as measured for the congruent stimuli in Figure 10.

Chapter 4: Summary and Conclusions

The present study showed that training listeners in all three conditions yielded improvement in all three conditions from pre-test to post-test. Although these findings confirm the value of training, they raise questions about the measurement of changes in audio-visual integration.

Results of testing revealed inconsistent patterns of integration. For the percent correct stimuli, integration as defined for the present study decreased with training from pre-test to post-test for three of the five listeners. For the other two listeners, integration of percent correct stimuli decreased from mid-test to post-test. In the present study, integration was defined as the difference between audio-visual performance and the better of the two single modalities, audio-only or visual only. The apparent decrease in integration likely reflects the greater degree of improvement in the A-only and V-only conditions, compared to A+V. Integration of McGurk-type discrepant stimuli, on the other hand, increased with training. This increase in integration was demonstrated by a greater percentage of fusion and combination responses with training. The apparent inconsistency of these two integration measures could be interpreted as arguing for the inclusion of discrepant stimuli in aural rehabilitation programs aimed as increasing integration. The fact that Ranta (2010) showed improvements in integration performance after training in only the audio-visual condition, together with the present findings that integration decreased after training in all three modalities, might further suggest that an aural rehabilitation program aimed as optimizing integration should provide far more training in the audio-visual conditions then in either single modality. However, such a conclusion might be superficial, based only on a single measure of integration.

Results of the present study indicate a need to evaluate various ways researchers have proposed to measure integration. One measure is integration efficiency (Tye-Murray et al., 2007), which calculates integration by looking at the predicted versus observed performance in the audio-visual condition. Oftentimes the predicted value is less than the observed value. For this reason, integration efficiency may be described as the amount that the listener benefits in audio-visual speech recognition beyond what is already measured in the auditory-only or visual-only performance. This alternative measure of integration can be used by professionals to quantify a patient's integration ability, as would be beneficial in an aural rehabilitation program for hearing impaired persons. The present study performed a preliminary look at the data using integration efficiency. No greater insights were found, but a more thorough look into this measure may prove beneficial.

Independent of how the measurement is made, it may be helpful to look further into the actual process of integration. One model proposed by Braida, the prelabeling (PRE) model of integration, suggests that integration of the auditory and visual stimuli occurs early. As listeners are presented with audio and visual stimuli, they subconsciously integrate the two modalities, resulting in a final decision regarding the stimuli presented. In contrast, the fuzzy logic model of perception (FLMP), proposed by Massaro and Cohen, attempts to attain the best fit of the obtained bimodal scores rather than predict the best speech performance. It is a process that integrates the single modalities later than the PRE model. When listeners are presented with audio and visual stimuli, they formulate decisions regarding the independent stimuli and then integrate the two decisions. In a study by Grant (2002), the PRE model was used because it was a better fit to estimate integration efficiency. It often equaled or exceeded the observed audio-visual performance, in contrast to the FLMP, which was equally likely to over-predict as it was

to under-predict the observed performance. The tendency of the PRE model to over-predict audio-visual results, instead of giving inconsistent results or under-predicting performance, may make it a more desirable model for integration.

Several other approaches could be taken to analyze the inconsistency of integration in the present study. It is possible that there are two different types of integration occurring in this study, one with congruent stimuli and another with discrepant stimuli. In order to further examine this notion, imagining techniques, such as fMRI, might be employed to further investigate where along the neural pathway these two types of integration are occurring. Finally, a new measure of integration could be designed for implementation in an aural rehabilitation program.

It is important to identify an accurate and reliable measure of integration so that it can be implemented in aural rehabilitation programs for hearing impaired individuals. Without an accurate and reliable measure of integration, it will be difficult for professionals to gauge how much benefit the aural rehabilitation program will provide hearing-impaired persons in daily living situations where they must integrate the auditory and visual signals.

Chapter 5: References

DiStefano, S. (2010). *Can audio-visual integration improve with training?* Senior Honors Thesis, The Ohio State University.

Feleppelle, N.M. (2008). The *role of the auditory signal in auditory-visual integration.* Audiology Capstone Project, The Ohio State University.

Gariety, M. (2009). *Effects of training on intelligibility and integration of sine-wave speech.* Senior Honors Thesis, The Ohio State University.

Grant, K.W. (2002) Measures of auditory-visual integration for speech understanding: A theoretical perspective (L). *The Journal of the Acoustical Society of America, 112 (1), 30-33.*

Grant, K.W., & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104 (4), 2438-2450.*

Jackson, P.L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review, 90(5),* 99-114.

James, K. (2009). *The effects of training on intelligibility of reduced information speech stimuli.* Senior Honors Thesis, The Ohio State University.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264, 746-748.*

Munhall, K.G., Kroos, C., Jozan, C., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception and Psychophysics,*

66, 574-583.

Ranta, A. (2010). *How does feedback impact training in audio-visual speech perception?* Senior Honors Thesis. The Ohio State University.

Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional cues. *Science, 212 (4497),* 947-950.

Shannon, R.V., Seng, F.G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270, 303-304.*

Shannon, R.V., F.G., Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *The Journal of the Acoustical Society of America*, *104 (4),* 2467-2475.

Small, L.H. (2004). *Fundamentals of Phonetics: A Practical Guide for Students (2nd Edition)*. Boston: Pearson.

Tye-Murray, N., Sommers, M.S., & Spehar, B. (2007). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing*, 28, 656-668.

Closed Set Response Sheet for Testing and Training

bat

pat

mat

dat

tat

nat

sat

zat

gat

kat

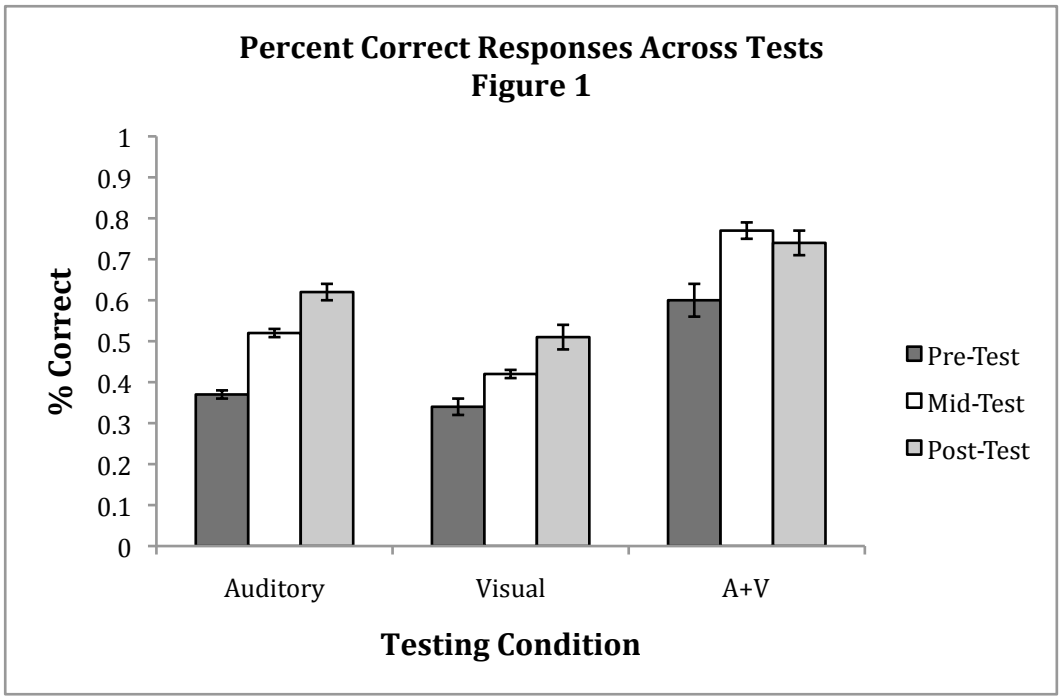
bgat

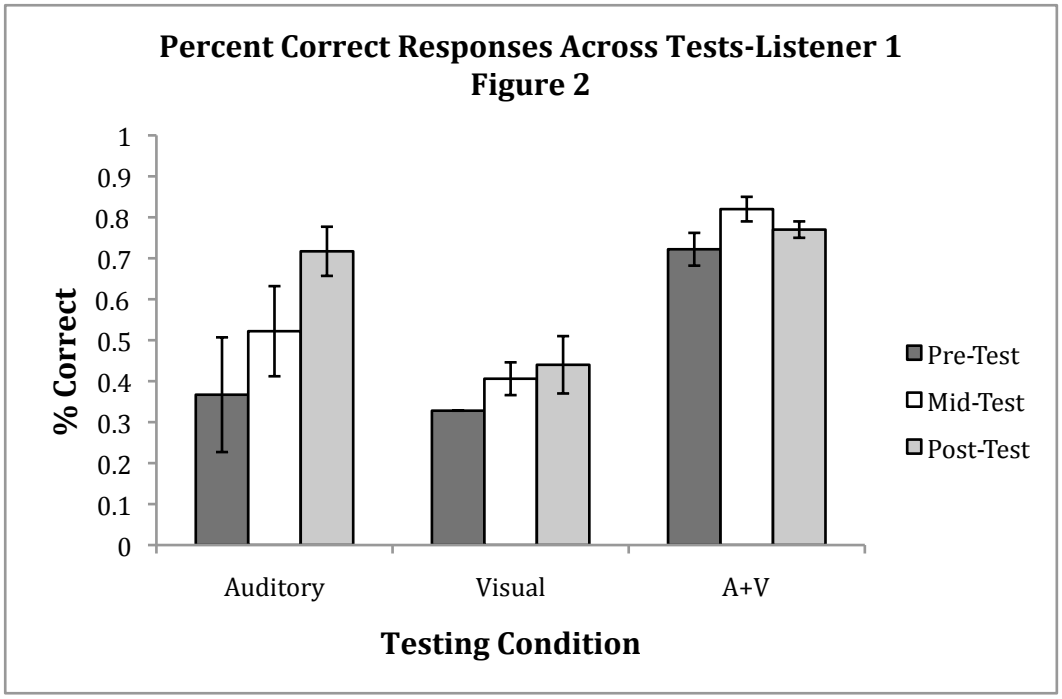
pkat

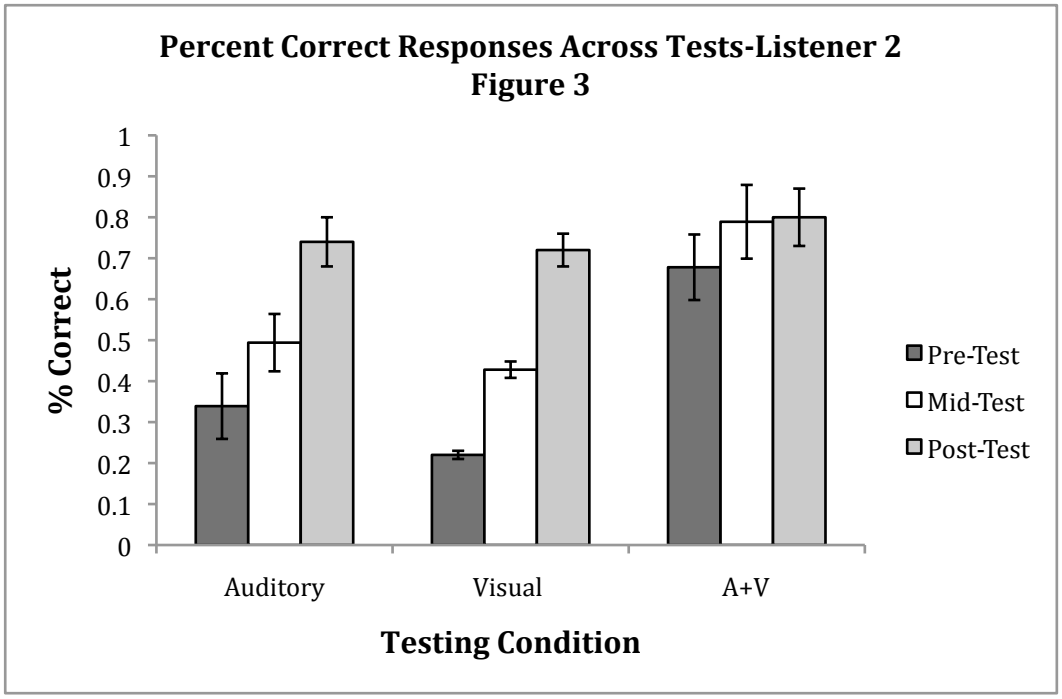
ptat

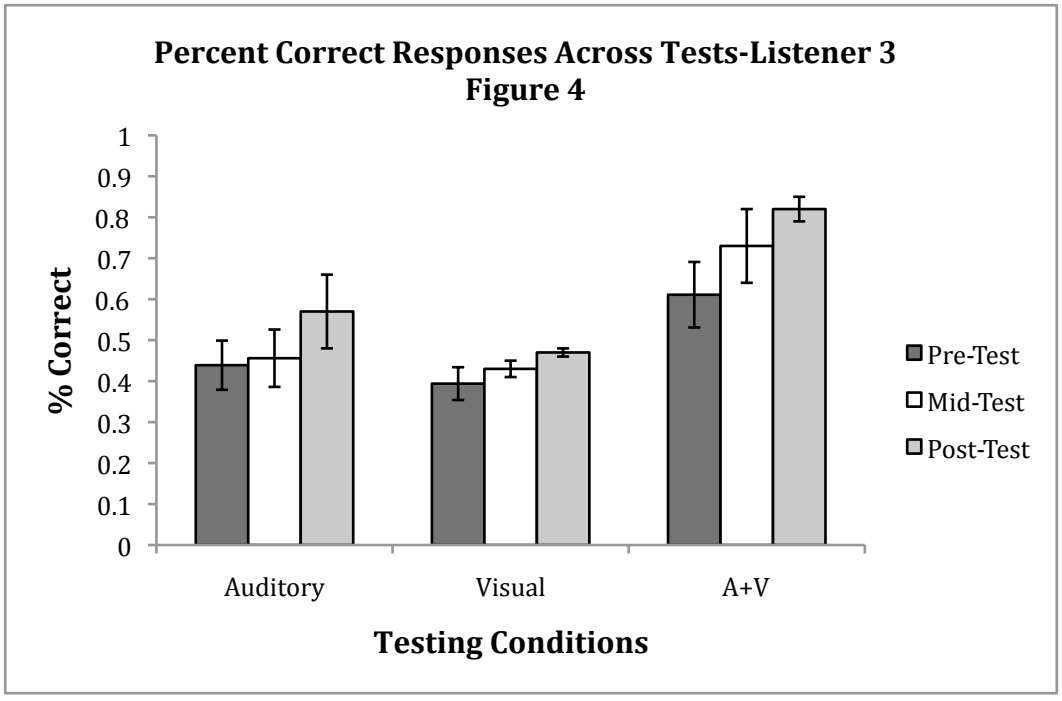
bdat

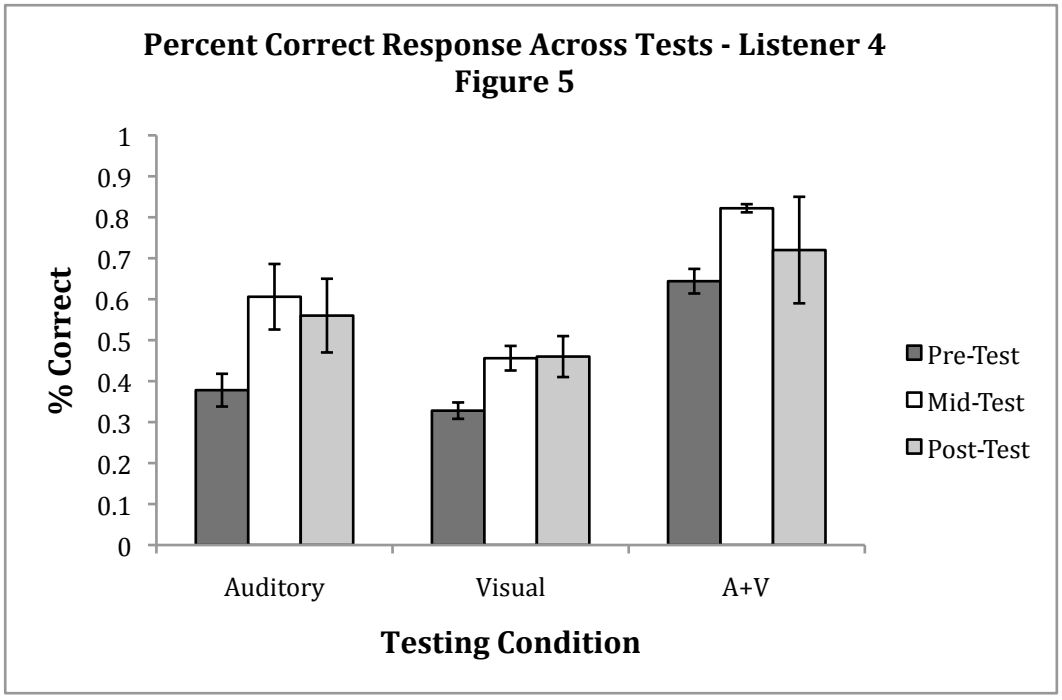
List of Figures

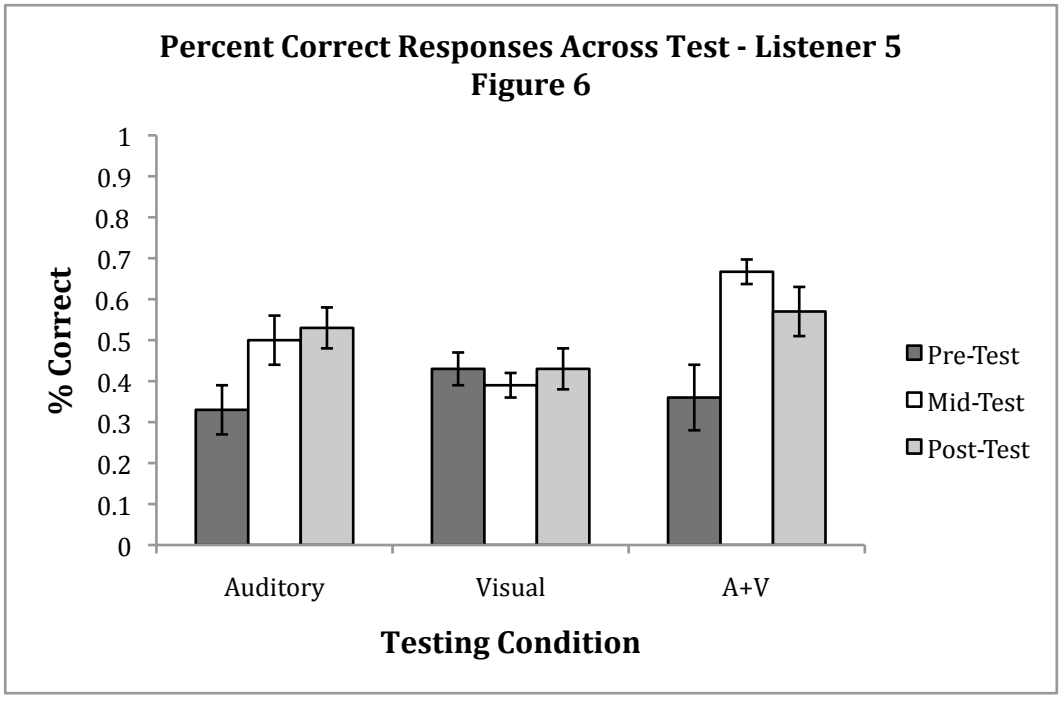

Percent Correct Responses Across Tests
Figure 1
% Correct
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Auditory
Visual
A+V
Testing Condition
Pre-Test
Mid-Test
Post-Test


Percent Correct Responses Across Tests-Listener 1
Figure 2
% Correct
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Auditory
Visual
A+V
Testing Condition
Pre-Test
Mid-Test
Post-Test

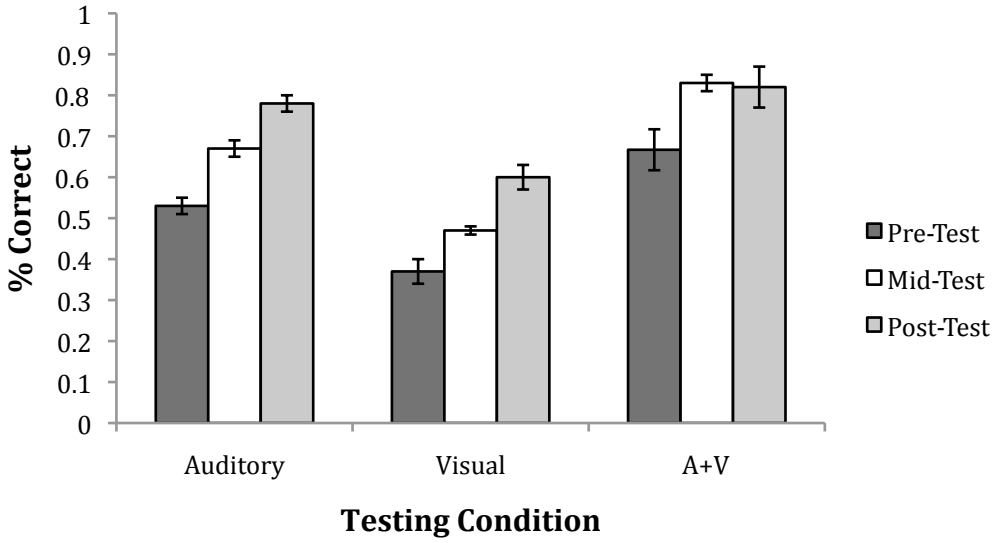

Percent Correct Responses Across Tests-Listener 2
Figure 3
% Correct
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Auditory
Visual
A+V
Testing Condition
Pre-Test
Mid-Test
Post-Test


Percent Correct Responses Across Tests-Listener 3
Figure 4
% Correct
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Auditory
Visual
A+V
Testing Conditions
Pre-Test
Mid-Test
Post-Test

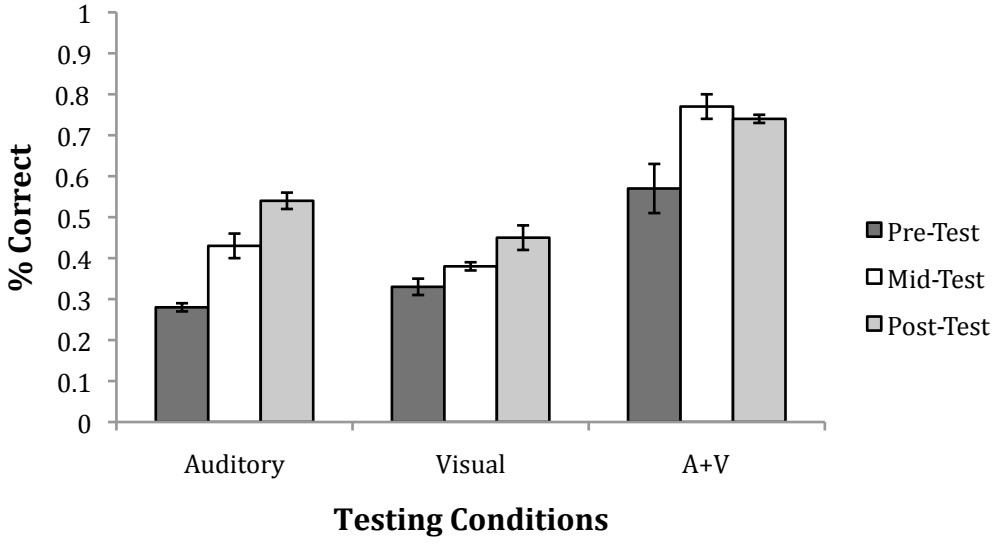

Percent Correct Response Across Tests - Listener 4
Figure 5
% Correct
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Auditory
Visual
A+V
Testing Condition
Pre-Test
Mid-Test
Post-Test

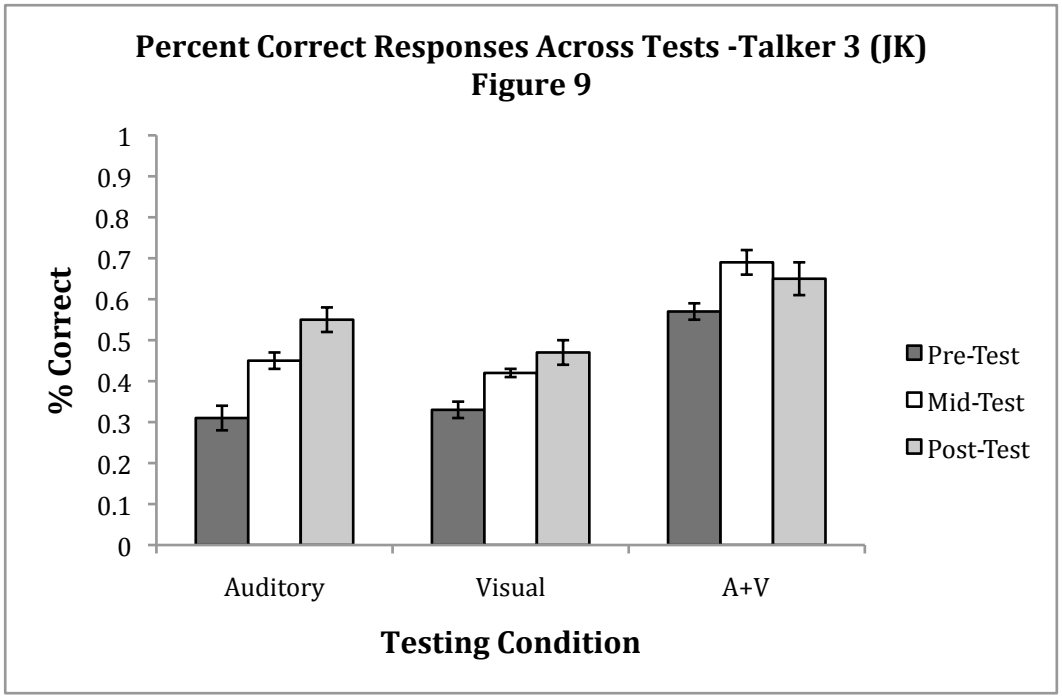

Percent Correct Responses Across Test - Listener 5
Figure 6
% Correct
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Auditory
Visual
A+V
Testing Condition
Pre-Test
Mid-Test
Post-Test

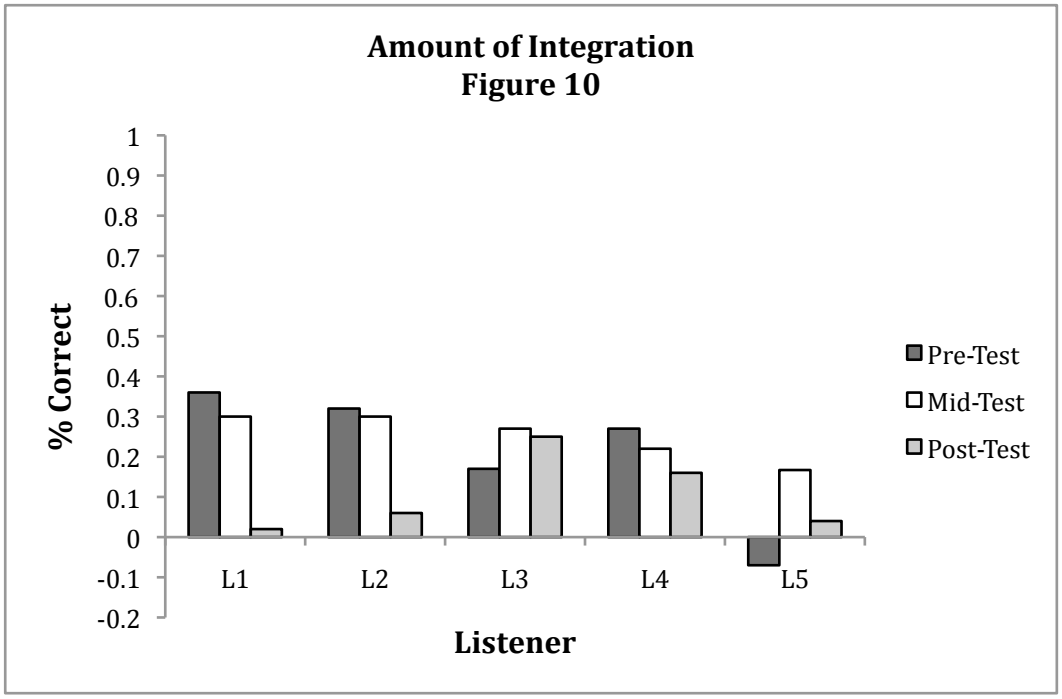
Percent Correct Responses Across Test -Talker 1 (LG)
Figure 7

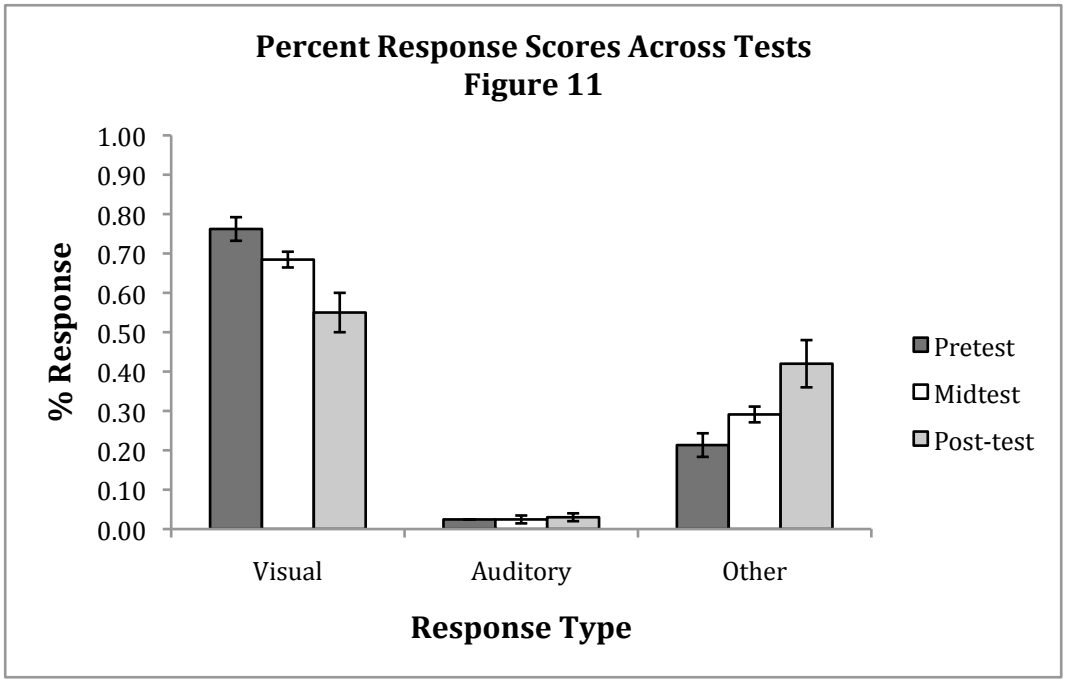


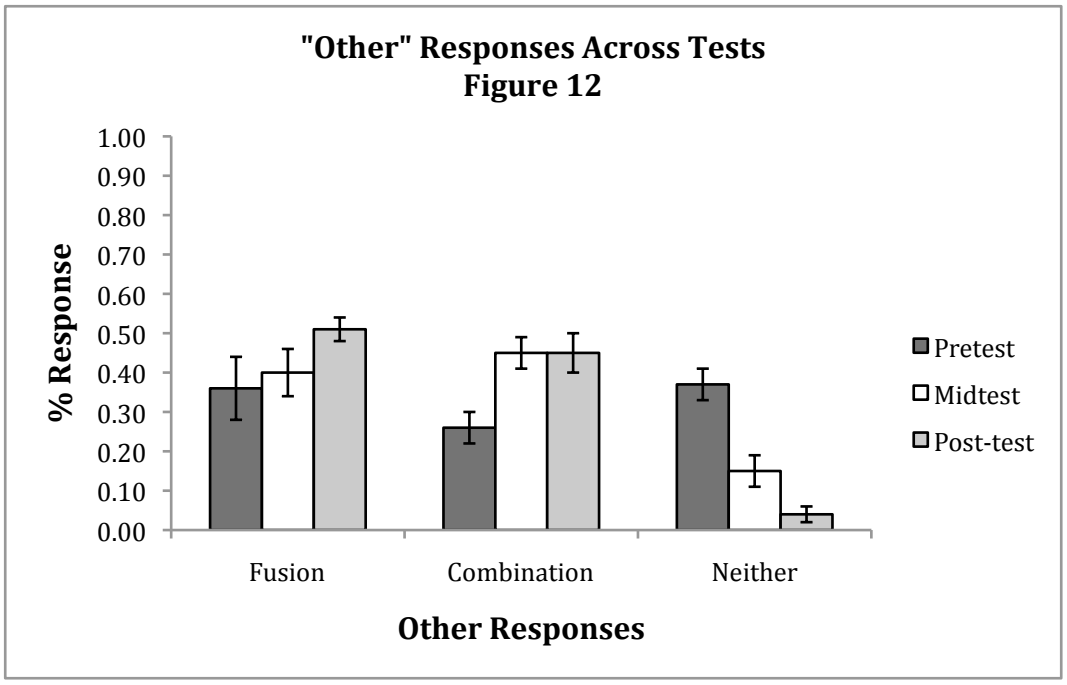
Percent Correct Responses Across Tests - Talker 2 (EA)
Figure 8




Percent Correct Responses Across Tests -Talker 3 (JK)
Figure 9
% Correct
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Auditory
Visual
A+V
Testing Condition
Pre-Test
Mid-Test
Post-Test


Amount of Integration
Figure 10
% Correct
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
L1
L2
L3
L4
L5
Listener
Pre-Test
Mid-Test
Post-Test


Percent Response Scores Across Tests
Figure 11
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
% Response
Visual
Auditory
Other
Response Type
Pretest
Midtest
Post-test


"Other" Responses Across Tests
Figure 12
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
% Response
Fusion
Combination
Neither
Other Responses
Pretest
Midtest
Post-test